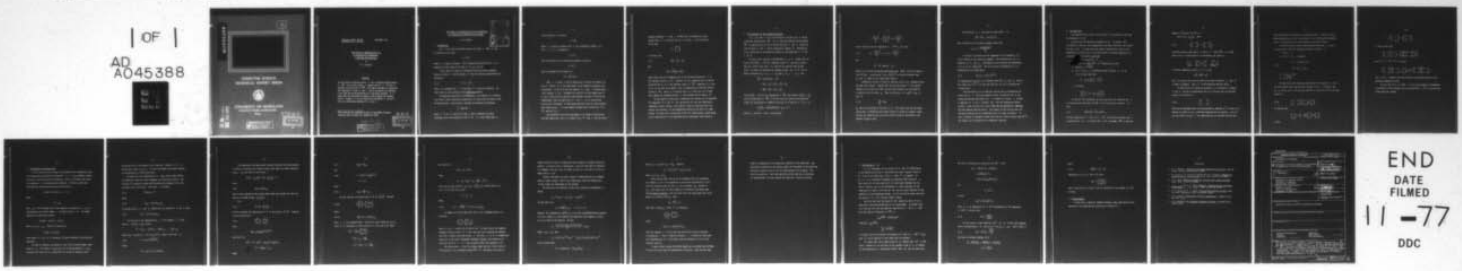
AD-A045 388 MARYLAND UNIV COLLEGE PARK DEPT OF COMPUTER SCIENCE F/G 12/1
THE EFFECTS OF ROUNDING ERROR ON AN ALGORITHM FOR DOWNDATING A --ETC(U)
SEP 77 G W STEWART N00014-76-C-0391
UNCLASSIFIED TR-582 NL



END
DATE FILMED
11-77
DDC

12
B.S.

COMPUTER SCIENCE
TECHNICAL REPORT SERIES

UNIVERSITY OF MARYLAND
COLLEGE PARK, MARYLAND
20742

Technical Report TR-582
ONR-N00014-76-C-0391-582

September 1977

The Effects of Rounding Error on an Algorithm for Downdating a Cholesky Factorization

by

G. W. Stewart*

Abstract

Let the positive definite matrix A have a Cholesky factorization $A = R^TR$. For a given vector x suppose that $\tilde{A} = A - xx^T$ has a Cholesky factorization $\tilde{A} = \tilde{R}^T\tilde{R}$. This paper considers an algorithm for computing $\tilde{R}$ from R and x and an extension for removing a row from the QR factorization of a regression problem. It is shown that the algorithm is stable in the presence of rounding errors. However, it is also shown that the matrix $\tilde{R}$ can be a very ill-conditioned function of R and x.

*This research was supported in part by the Office of Naval Research under Contract No. N00014-76-C-0391.

The Effects of Rounding Error on an Algorithm for Downdating a Cholesky Factorization

G. W. Stewart

1. Introduction

Let A be a positive definite matrix of order p. Then A can be factored in the form

$$A = R^T R$$

where R is upper triangular. This "Cholesky factorization" of A is unique up to the signs of the rows of R (e.g. see [6]

In this paper we shall be concerned with the following problem. Given a p-vector x and the matrix R find the Cholesky factorization of the matrix

$$\tilde{A} = A - xx^T , \tag{1.1}$$

where it is assumed that x is such that $\tilde{A}$ is positive definite. We shall refer to this problem as the downdating problem.

An important application of downdating is the removal of an observation from a linear regression problem that is being solved by means of the QR factorization. Specifically, consider the problem of minimizing

$$\rho^2 = \|X\beta - y\|^2 ,$$

where X is an $n \times p$ matrix of rank p and $\|\cdot\|$ denotes the usual Euclidean vector norm defined by $\|x\|^2 = x^T x$. It is well known that X

can be factored in the form

$$X = QR ,$$

where R is upper triangular and Q has orthonormal columns, i.e. $Q^TQ = I$. If z is defined by

$$z = Q^Tr ,$$

then the solution of the regression problem is given by

$$\beta = R^{-1}z$$

and the residual sum of squares by

$$\rho^2 = \|y\|^2 - \|z\|^2 .$$

When n is large, it may be impossible to retain the elements of the $n \times p$ matrix Q in the main memory of the computer performing the calculation. In this case one may compute R, z, and ρ without explicitly forming Q [3,5]. Although this suffices for the computation of β, one is left with the problem of performing a variety of statistical computations when one knows only R, z, and ρ. (It is interesting to note that aficiandos of the normal equations have the same problem; they cannot retain X in main memory and must work instead with X^TX, X^Ty, and $\|y\|^2$ [1].)

One frequently occurring requirement is to remove an observation from the regression, that is to remove a row x^T from X and the corre-

sponding component η from y. Without loss of generality we may suppose that x^T is the last row of X, so that X can be written in the form

$$X = \begin{pmatrix} \tilde{X} \\ x \end{pmatrix} .$$

It follows that

$$\tilde{X}^T\tilde{X} = X^TX - xx^T . \tag{1.2}$$

Now

$$X^TX = R^TQ^TQR = R^TR ,$$

which shows that the triangular part of the QR factorization of X is the Cholesky factor of X^TX. Likewise $\tilde{R}$, the triangular part of the QR factorization of $\tilde{X}$, is the Cholesky factor of $\tilde{X}^T\tilde{X}$. Comparing (1.1) and (1.2), we see that the problem is one of downdating the Cholesky factorization of X^TX. There is, of course, more to it than this, for we must also compute the downdated vector $\tilde{z}$ and residual sum of squares $\tilde{\rho}^2$.

In this paper we shall give a rounding error analysis of an algorithm for computing $\tilde{R}$, $\tilde{z}$, and $\tilde{\rho}$. Our conclusions are that the algorithm is remarkably stable; however, this stability does not guarantee that the results are accurate, for the downdating problem can be quite ill conditioned. We begin with a discussion of this ill-conditioning, before going on to a description of the algorithm and the subsequent error analysis.

2. The condition of the downdating problem

Let A, $\tilde{A}$, and x be as in the previous section with A having a Cholesky factorization R^TR. Let $\tilde{A}$ have the Cholesky factorization $\tilde{R}^T\tilde{R}$. In applications we will of course not know A and B. Rather we will be given R and x and be required to compute $\tilde{R}$. Consequently, we are interested in assessing the effects of perturbations in R and x on $\tilde{R}$.

We shall first consider a perturbation E in R. Assume that the matrix $(R+E)^T(R+E) - xx^T$ has a Cholesky factor $\bar{R}$. We wish to assess the size $\|\tilde{R}-\bar{R}\|$, where here $\|\cdot\|$ denotes the spectral matrix norm [6,7]. We begin by comparing the singular values [6] of $\tilde{R}$ and $\bar{R}$, which we denote by $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \ldots \geq \tilde{\sigma}_p$ and $\bar{\sigma}_1 \geq \bar{\sigma}_2 \geq \ldots \geq \bar{\sigma}_p$. Now

$$\begin{aligned}\bar{R}^T\bar{R} &= (R+E)^T(R+E) - xx^T \\ &= R^TR - xx^T + R^TE + E^TR + E^TE \\ &= \tilde{R}^T\tilde{R} + R^TE + E^TR + E^TE \ .\end{aligned}$$

Since $\tilde{\sigma}_1^2,\tilde{\sigma}_2^2,\ldots,\tilde{\sigma}_p^2$ are the eigenvalues of $\tilde{R}^T\tilde{R}$ and likewise $\bar{\sigma}_1^2,\bar{\sigma}_2^2,\ldots,\bar{\sigma}_p^2$ are the eigenvalues of $\bar{R}^T\bar{R}$, it follows from the classical perturbation theory for eigenvalues of symmetric matrices [6,7] that for $i = 1,2,\ldots,p$

$$|\tilde{\sigma}_i^2-\bar{\sigma}_i^2| \leq \|R^TE+E^TR+E^TE\| \leq 2\sigma_1\varepsilon + \varepsilon^2 \ ,$$

where $\sigma_1 = \|R\|$ and $\varepsilon = \|E\|$. In particular

$$1 - \frac{2\sigma_1\varepsilon+\varepsilon^2}{\tilde{\sigma}_i^2} \le \left(\frac{\bar{\sigma}_i}{\tilde{\sigma}_i}\right)^2 \le 1 + \frac{2\sigma_1\varepsilon+\varepsilon^2}{\tilde{\sigma}_i^2} ,$$

and it follows from the inequality $|1 - \sqrt{1 + x}| \le |x|$ that

$$\tilde{\sigma}_i - \frac{2\sigma_1\varepsilon+\varepsilon^2}{\tilde{\sigma}_i^2} \le \bar{\sigma}_i \le \tilde{\sigma}_i + \frac{2\sigma_1\varepsilon+\varepsilon^2}{\tilde{\sigma}_i^2} . \tag{2.1}$$

Now

$$\|\tilde{R} - \bar{R}\| \ge \max|\tilde{\sigma}_i - \bar{\sigma}_i| ;$$

hence (2.1) has the disturbing implication that $\|\tilde{R}-\bar{R}\|$ can be as large as $(2\sigma_1\varepsilon+\varepsilon^2)/\tilde{\sigma}_p$. In particular if $\tilde{\sigma}_p \le \sqrt{2\sigma_1\varepsilon+\varepsilon^2}$, we cannot guarantee that $\tilde{\sigma}_p$ and $\bar{\sigma}_p$ agree in any significant figures.

Casting the results in terms of relative errors (e.g. rounding errors) may make this clearer. Suppose that the original matrix R has nonzero elements all of about the same size, and these elements are perturbed by a relative error of order ε_M. Then in the above, $\varepsilon \cong \varepsilon_M\|R\| = \varepsilon_M\sigma_1$, so that if

$$\frac{\sigma_1}{\tilde{\sigma}_p} \le \sqrt{\varepsilon_M} \tag{2.2}$$

$\tilde{\sigma}_p$ may be obliterated by the error in R. The square root has the implication that in downdating one cannot tolerate a spread of singular values of half the computational precision without losing all precision in the smallest singular value.

Perturbations in x have much the same effect. If

$$\bar{R}^T\bar{R} = R^TR - (x+f)(x+f)^T,$$

then a repetition of the above argument shows that

$$|\sigma_i - \bar{\sigma}_i| \le \frac{2\|x\|\|f\| + \|f\|^2}{\tilde{\sigma}_i^2}.$$

It should be observed that the dependence of the bounds on $\tilde{\sigma}_i^{-2}$ can be removed by the following argument. The derivation of (2.1) is symmetric in $\tilde{\sigma}_i$ and $\bar{\sigma}_i$. Consequently we may replace the denominator in (2.1) by $\mu_i^2 = \max\{\tilde{\sigma}_i^2, \bar{\sigma}_i^2\}$. But we also have the bound $|\tilde{\sigma}_i - \bar{\sigma}_i| \le \mu_i$. Combining the two bounds gives

$$|\tilde{\sigma}_i - \bar{\sigma}_i| \le (2\sigma_1\varepsilon + \varepsilon^2)^{1/3}. \tag{2.3}$$

In computational practice it is unlikely that both $\tilde{\sigma}_i$ and $\bar{\sigma}_i$ will be less than $(2\sigma_1\varepsilon+\varepsilon^2)^{1/2}$, so that the cube root in (2.3) is effectively a square root.

That the bound (2.1) is realistic can be seen by considering the scalar case $p = 1$. This case actually arises in practice; for in the regression problem mentioned in §1, X becomes an n-vector, x becomes a component of X, and R becomes $\|X\|$. Thus the downdating problem becomes: given the norm of a vector find the new norm after a component has been removed from the vector. The results of this section have the following implications for downdating norms in t-digit arithmetic. If ever a sequence of downdates reduces the norm by a factor greater than $10^{t/2}$, the results can be expected to be completely spurious.

3. The algorithm

The algorithm described in this section is an extension of one that has appeared in [2,5].

We shall use the notation introduced in §1. We assume that the reader is familiar with computations with plane rotations (for details see [6] or [7]). In order that the several computational steps of the algorithm will not be lost in the derivations, we proceed immediately to a description of the entire algorithm.

1. Solve the system $a^T R = x^T$.
2. If $\|a\| \geq 1$, report $R^T R - xx^T$ indefinite and stop.
3. Compute $\alpha = \sqrt{1-\|a\|^2}$.
4. For $i = p,p-1,\ldots,1$ determine plane rotations U_i in the $(i,p+1)$ plane such that

$$U_1 \ldots U_{p-1} U_p \begin{pmatrix} a \\ \alpha \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

5. Calculate

$$\begin{pmatrix} \tilde{R} \\ x^T \end{pmatrix} = U_1 \ldots U_{p-1} U_p \begin{pmatrix} R \\ 0 \end{pmatrix}.$$

To justify this algorithm, we first show that the condition $\|a\| < 1$ is necessary and sufficient for $R^T R - xx^T$ to be positive definite. In fact

$$(R^T R - xx^T) = R^T (I - aa^T) R \,. \tag{3.1}$$

Now the eigenvalues of $I-aa^T$ are $1 - \|a\|^2$ of multiplicity unity and 1 of multiplicity $p-1$. It follows that $I-aa^T$, and hence $\tilde{R}^T\tilde{R}$, is positive

definite if and only if $\|a\|^2 < 1$.

Let $Q = U_1 \ldots U_{p-1} U_p$. Then

$$Q \begin{pmatrix} a & R \\ \alpha & 0 \end{pmatrix} = \begin{pmatrix} o & \tilde{R} \\ \beta & b^T \end{pmatrix}, \tag{3.2}$$

in which we must verify that $\beta = 1$ and $b = x$. Since $Q^T Q = I$, if each side of (3.2) is multiplied by its transpose, the result is

$$\begin{pmatrix} 1 & a^T R \\ R^T a & R^T R \end{pmatrix} = \begin{pmatrix} \beta^2 & \beta b^T \\ \beta b & \tilde{R}^T \tilde{R} + b b^T \end{pmatrix}.$$

It follows immediately that $\beta = 1$, $b = x$, and

$$R^T R = \tilde{R}^T \tilde{R} + x x^T .$$

But it is easily seen from the form of the plane rotations U_i that $\tilde{R}$ is upper triangular. Hence $\tilde{R}$ is the downdated Cholesky factor.

In applications to regression problems, it is necessary to compute $\tilde{z}$ and $\tilde{\rho}$. One way of approaching this is to observe that the Cholesky factor of $(X,y)^T(X,y)$ is

$$\begin{pmatrix} R & z \\ 0 & \rho \end{pmatrix} . \tag{3.3}$$

Thus the new decomposition can be determined by removing $(x^T \; \eta)$ from (3.3). However we prefer to use a different algorithm for two reasons. First if one has several vectors y, the algorithm must be repeated for each one,

with considerable diseconomies in time and storage. Second, the augmented downdating may fail, even though R by itself can be downdated, and it is desirable not to confound these sources of failure.

In the description of our proposed algorithm below, c_i and s_i are the cosines and sines defining the plane rotations U_i.

1. Set $\tilde{\eta}_0 = \eta$
2. For $i = 1,2,\ldots,p$ compute

$$\tilde{z}_i = (z_i + s_i\tilde{\eta}_{i-1})/c_i$$

(3.4)

$$\tilde{\eta}_i = s_i\tilde{z}_i + c_i\tilde{\eta}_{i-1}$$

3. If $\tilde{\eta}_p > \rho$ stop
4. $\tilde{\rho} = \sqrt{\rho^2 - \tilde{\eta}_p^2}$

To show that these formulas indeed produce the required $\tilde{z}$ and $\tilde{\rho}$, we first observe that they are well defined, since $\alpha \neq 0$ implies that no c_i can be zero. Now the two relations in step two of (3.4) are equivalent to

$$\begin{pmatrix} c_i & -s_i \\ s_i & c_i \end{pmatrix} \begin{pmatrix} \tilde{z}_i \\ \tilde{\eta}_{i-1} \end{pmatrix} = \begin{pmatrix} z_i \\ \tilde{\eta}_i \end{pmatrix} . \tag{3.5}$$

It follows that

$$\begin{pmatrix} z \\ \tilde{\eta}_p \end{pmatrix} = U_p^T \ldots U_2^T U_i^T \begin{pmatrix} \tilde{z} \\ \eta \end{pmatrix} = Q^T \begin{pmatrix} \tilde{z} \\ \eta \end{pmatrix},$$

whence

$$Q\begin{pmatrix} R & z \\ 0 & \tilde{\eta}_p \end{pmatrix} = \begin{pmatrix} \tilde{R} & \tilde{z} \\ x^T & \eta \end{pmatrix} .$$

It then follows that

$$\begin{pmatrix} R & z \\ 0 & \rho \end{pmatrix}^T \begin{pmatrix} R & z \\ 0 & \rho \end{pmatrix} = \begin{pmatrix} \tilde{R}^T\tilde{R}+xx^T & \tilde{R}^T\tilde{z}+\eta x \\ \tilde{z}^T\tilde{R}+\eta x^T & z^Tz+\rho^2 \end{pmatrix} .$$

But $z^Tz + \rho^2 = \tilde{z}^T\tilde{z} + \tilde{\rho}^2 + \eta^2$; hence

$$\begin{pmatrix} R & z \\ 0 & \rho \end{pmatrix}^T \begin{pmatrix} R & z \\ 0 & \rho \end{pmatrix} - \begin{pmatrix} x \\ \eta \end{pmatrix}(x\ \eta)^T = \begin{pmatrix} \tilde{R} & \tilde{z} \\ 0 & \tilde{\rho} \end{pmatrix}^T \begin{pmatrix} \tilde{R} & \tilde{z} \\ 0 & \tilde{\rho} \end{pmatrix} ,$$

and $\tilde{z}$ and $\tilde{\rho}$ comprise the last column of the Cholesky factor of the downdated augmented system.

We note that if $\eta_p > \rho$, then ρ is not large enough to accommodate the decrease in the residual due to the deletion of (x^T,η), and the algorithm should be stopped.

4. The effects of rounding error

In this section we shall adopt the conventions and assumptions usual in floating-point rounding-error analyses. If e is an arithmetic expression with a specified order of evaluation, $fl(e)$ will denote the result of evaluating e in floating-point arithmetic. We shall assume that floating-point multiplication and division satisfy

$$fl(a \circ b) = a \circ b(1+\varepsilon), \quad \circ = \times, \div,$$

where

$$|\varepsilon| \leq \varepsilon_M .$$

Here ε_M is the rounding unit of the computer in question (i.e. ε_M is approximately the largest number ε for which $fl(1+\varepsilon) = 1$). We assume addition and subtraction satisfy

$$fl(a \pm b) = a(1+\varepsilon_1) \pm b(1+\varepsilon_2)$$

where $|\varepsilon_1|, |\varepsilon_2| \leq \varepsilon_M$. Finally we assume that

$$fl(\sqrt{a}) = (1+\varepsilon)\sqrt{a} ,$$

where again $|\varepsilon| \leq \varepsilon_M$. As is customary, we ignore problems of overflow and underflow.

In order to simplify our bounds we shall freely discard higher order terms in ε_M. For example $(1+\varepsilon_M)(1+\varepsilon_M)$ will be approximated by $1 + 2\varepsilon_M$. Although our results will no longer have the status of theorems, their

derivation will be considerably less cluttered. Moreover, if p is sufficiently small, say $p\varepsilon_M < .01$, then the bounds can be made rigorous by multiplying by a factor near unity.

We begin with the computation of a. Here, and in what follows, all quantities stand for their computed, not their true, values. The solution of triangular systems has been analyzed elsewhere [6,7], and we merely quote the results. The vector a satisfies

$$a^T(R+F) = x^T \tag{4.1}$$

where

$$|f_{ij}| \leq (j+2)|r_{ij}|\varepsilon_M . \tag{4.2}$$

It follows that if r_j and f_j denote the j-th columns of R and F, then

$$\|f_j\| \lesssim \sqrt{j}(j+2)\|r_j\|\varepsilon_M . \tag{4.3}$$

We turn now to the computation of α. If we compute α^2 in the order $1 - (a_1^2+a_2^2+\ldots+a_p^2)$ we have

$$\alpha^2 = (1+\varepsilon_0) - a_1^2(1+\varepsilon_1) - a_2^2(1+\varepsilon_2) - \ldots - a_p^2(1+\varepsilon_p)$$

where $|\varepsilon_0| \leq \varepsilon_M$ and $|\varepsilon_i| \leq (p-i+3)\varepsilon_M$ $(i\geq 1)$. Hence, since $\|a\|^2 < 1$,

$$\alpha = (1+\tau_1)\sqrt{1-\|a\|^2+\tau_2} \tag{4.4}$$

where

$$|\tau_1| \lesssim \varepsilon_M, \quad |\tau_2| \lesssim (p+3)\varepsilon_M .$$

The computation and application of plane rotations has been analyzed in detail by Wilkinson [6], where he shows that there are exact rotations $\hat{U}_1,\hat{U}_2,\ldots,\hat{U}_p$ such that for any vector v

$$fl(U_1\ldots U_{p-1}U_p v) = \hat{U}_1\ldots\hat{U}_{p-1}\hat{U}_p v + g$$

where

$$\|g\| \lesssim 6p\|v\|\varepsilon_M \,. \tag{4.4}$$

Here we have suppressed some second order terms that account for the slow growth in a bound on $\|U_1\ldots U_{p-1}U_p v\|$.

Let

$$\hat{Q} = \hat{U}_1\ldots\hat{U}_{p-1}\hat{U}_p \,. \tag{4.5}$$

We first consider the application of $\hat{Q}$ to the vector $(a^T,\alpha)^T$. From the results quoted above

$$\hat{Q}\begin{pmatrix} a \\ \alpha \end{pmatrix} = \begin{pmatrix} g_0 \\ \beta \end{pmatrix} \tag{4.6}$$

where

$$\|g_0\| \lesssim 6p\varepsilon_M(\|a\|^2+\alpha^2)^{1/2} \,.$$

Now from (4.4)

$$\|a\|^2 + \alpha^2 = \|a\|^2 + (1+\tau_1)^2(1-\|a\|^2+\tau_2)$$

$$\cong 1 + 2\|a\|\tau_1 + \tau_2 \,.$$

Hence

(4.7) $$\|g_0\| \lesssim 6p\varepsilon_M$$

and

(4.8) $$\beta = (\|a\|^2+\alpha^2-\|g_0\|^2)^{1/2}$$

$$= 1 + \sigma_0 ,$$

where

(4.9) $$|\sigma_0| \lesssim \frac{p+5}{2}\ \varepsilon_M .$$

We next consider the application of $\hat{Q}$ to $(r_j^T,0)^T$. We have

(4.10) $$\hat{Q}\begin{pmatrix} r_j^T \\ 0 \end{pmatrix} = \begin{pmatrix} \tilde{r}_j+g_j \\ \xi_j+\gamma_j \end{pmatrix} ,$$

where

(4.11) $$\|g_j\|, |\gamma_j| \lesssim 6p\|r_j\|\varepsilon_M .$$

Here ξ_j is the computed value. We wish to find a bound on $|x_j-\xi_j|$. Since $\hat{Q}$ is orthogonal, we have from (4.6), (4.8) and (4.10) that

$$a^T r_j = (g_0^T,\beta)\begin{pmatrix} \tilde{r}_j+g_j \\ \xi_j+\gamma_j \end{pmatrix}$$

$$\tilde{=} \beta\xi_j + \gamma_j + g_0^T r_j$$

$$= \xi_j + \sigma_0\xi_j + \gamma_j + g_0^T\tilde{r}_j .$$

But from (4.1)

$$a^T r_j = x_j + a^T f_j \ .$$

Hence

$$x_j - \xi_j \cong \sigma_0 \xi_j + \gamma_j + g_0^T \tilde{r}_j - a^T f_j \ .$$

Since up to terms of order ϵ_M, $\|r_j\| = \|(\tilde{r}_j^T, \xi_j)\|$, we have from (4.3), (4.7), (4.9), and (4.11) that

$$x_j - \xi_j = \sigma_j$$

where

$$|\sigma_j| \lesssim \left[\frac{13p+5}{2} + \sqrt{j}\ (j+2) \right] \|r_j\| \epsilon_M \ . \tag{4.12}$$

To summarize we have shown that there is an orthogonal matrix $\hat{Q}$ such that

$$\hat{Q} \begin{pmatrix} R \\ 0 \end{pmatrix} = \begin{pmatrix} \tilde{R}+G \\ x^T+s^T \end{pmatrix} ,$$

where G and s satisfy (4.11) and (4.12). In other words, the computed downdated Cholesky factor $\tilde{R}$ is very near the factor obtained by downdating with a slightly perturbed vector x. The error G in $\tilde{R}$ is unimportant, except as it may affect subsequent downdates; however, the results of §2 show that the error s in x may seriously affect the accuracy of $\tilde{R}$.

Two other points. First, the higher order term in (4.12) is due to the solution of the triangular system $a^T R = x^T$. The factor $(j+2)$ can be

removed from this term by accumulating inner products in double precision; however, in practice this is unnecessary, since the term does not dominate its companion (and this only in column p) until p = 40, and it is not yet double when p = 150.

Second, the bounds are given column by column and hence are independent of column scaling. This is not surprising, since the computations in each column are independent of one another.

We turn now to the analysis of the errors involved in downdating $\tilde{z}$. Define

$$w_i = (z_1, z_2, \ldots, z_{i-1}, \tilde{z}_i, \ldots, \tilde{z}_p, \tilde{\eta}_i)^T ,$$

so that from (3.5)

$$w_i = fl(U_i^T w_{i-1}), \quad i = 1,2,\ldots,p .$$

However, the evaluation of $fl(U^T w_{i-1})$ is not the straightforward one implied by (3.5); rather it is the indirect one implied by the formulas in step 2 of (3.4), which we now analyze. We have

$$\tilde{z}_i = \frac{[z_i(1+\varepsilon_1)+s_i\tilde{\eta}_{i-1}(1+\varepsilon_2)(1+\varepsilon_3)]}{c_i}(1+\varepsilon_4)$$

where $|\varepsilon_i| \le \varepsilon_M$. Thus

$$z_i = c_i\tilde{z}_i(1+\varepsilon_1)^{-1}(1+\varepsilon_4)^{-1} - s_i\tilde{\eta}_{i-1}(1+\varepsilon_2)(1+\varepsilon_3)(1+\varepsilon_1)^{-1} ,$$

and it follows that

$$z_i = c_i\tilde{z}_i(1+\varepsilon_5) - s_i\tilde{\eta}_{i-1}(1+\varepsilon_6)$$

where $|\varepsilon_5| \lesssim 2\varepsilon_M$ and $|\varepsilon_6| \lesssim 3\varepsilon_M$. Likewise

$$\tilde{\eta}_i = s_i\tilde{z}_i(1+\varepsilon_7) + c_i\tilde{\eta}_{i-1}(1+\varepsilon_8) ,$$

where $|\varepsilon_7|, |\varepsilon_8| \lesssim 2\varepsilon_M$.

These results show that as far as rounding errors are concerned, the formulas in (3.4) are equivalent to the direct application of (3.5), with the exception that the term ε_6 has the bound $3\varepsilon_M$ instead of $2\varepsilon_M$. This means that the error analysis of Wilkinson cited above goes through mutatis mutandis, with the result that the right hand side of the bound (4.4) becomes $7p\|v\|\varepsilon_M$. Hence

$$w_p = Q^T w_0 + g ,$$

where $\|g\| \lesssim 7p\|w_0\|\varepsilon_M$. Since $\hat{Q}$ is orthogonal,

$$\hat{Q}\begin{pmatrix} z \\ \tilde{\eta} \end{pmatrix} = \begin{pmatrix} \tilde{z}+h \\ \eta+\tau \end{pmatrix} ,$$

where

$$\|h\|, |\tau| \lesssim 7p(\|\tilde{z}\|^2+\eta^2)\varepsilon_M .$$

Thus the computed $\tilde{z}$ is very near the vector that would be obtained by downdating z with a slightly perturbed η. It should be noted that the transformation $\hat{Q}$ is the same as the one defined by (4.5) in the previous analysis.

It goes without saying that these bounds are an extreme over-estimate of the errors that would be encountered in practice. None the less they

suffice to demonstrate the exceptional stability of the algorithm. Any inaccuracies observed in the results cannot be attributed to the algorithm; they must instead be due to the ill-conditioning of the problem. This raises the question: does the algorithm provide some way of detecting ill conditioning? We shall answer this question in the next section.

5. The meaning of $\|a\|$

It is a consequence of the results of §2 that ill conditioning in the downdating problem is associated with small singular values in $\tilde{R}$. In §3 it was shown that if $\|a\| = 1$, then $\tilde{R}$ is singular, i.e. $\tilde{\sigma}_p = 0$. It is therefore reasonable to conjecture that values of a near unity will be associated with ill-conditioned problems and vice versa. However, just as the determinant is a poor indicator of the condition of a matrix, the value of $\|a\|$ may be a poor indicator of the condition of the downdating problem. In this section we shall show that the value of $\|a\|$ will reliably signal trouble.

We first show that the value of $\|a\|$ cannot cry wolf; if it is near unity, then the problem must be ill conditioned. It follows from (3.1) and the fact that the smallest eigenvalue of $I - aa^T$ is $1 - \|a\|^2$ that the smallest eigenvalue of $\tilde{R}^T\tilde{R}$ is

$$\lambda_{min}(\tilde{R}^T\tilde{R}) = \|R\|_2^2(1-\|a\|^2) .$$

Since $\tilde{\sigma}_p^2 = \lambda_{min}(\tilde{R}^T\tilde{R})$,

$$\frac{\tilde{\sigma}_p}{\sigma_1} \leq \sqrt{1-\|a\|^2} .$$

It follows from the discussion surrounding (2.2) that if $1 - \|a\|^2 = O(\varepsilon_M)$ then $\tilde{R}$ can be expected to lose about half its accuracy.

We cannot show that a small value of $\tilde{\sigma}_p$ implies that $\|a\|$ is near unity. However we can show that if any singular value of R is reduced in the downdating by a significant factor, then $\|a\|$ must be near unity.

We start by developing an expression for $\|a\|^2$. First

$$\|a\|^2 = x^T R^{-1} R^{-T} x = x^T (R^T R)^{-1} x$$

$$= x^T (\tilde{R}^T \tilde{R} + xx^T)^{-1} x$$

$$= x^T \tilde{R}^{-1} (I + \tilde{R}^{-T} x x^T \tilde{R}^{-1}) \tilde{R}^{-T} x \ .$$

Set

$$b = \tilde{R}^{-T} x \ ,$$

so that

$$\|a\|^2 = b^T (I + bb^T)^{-1} b \ .$$

Since b is an eigenvector of $I + bb^T$ corresponding to the eigenvalue $1 + \|b\|^2$, it follows that

$$\|a\|^2 = \frac{\|b\|^2}{1+\|b\|^2} \ . \tag{5.1}$$

We next obtain a lower bound on $\|b\|^2$. Let $\tilde{v}_i$ be the right singular vector corresponding to $\tilde{\sigma}_i$ and let $\tilde{V}_i = (\tilde{v}_i, \tilde{v}_{i+1}, \ldots, \tilde{v}_p)$. Then if $\tilde{V}_i^T x \neq 0$

$$\|b\| = \|R^{-T} x\| \geq \frac{\|\tilde{V}_i^T x\|}{\tilde{\sigma}_i} \ . \tag{5.2}$$

But from the minimax theorems [6,7]

$$\sigma_i^2 \leq \|\tilde{V}_i^T R^T R \tilde{V}_i\| \leq \|\tilde{V}_i^T \tilde{R}^T \tilde{R} \tilde{V}_i\| + \|\tilde{V}_i^T x x^T \tilde{V}_i\|$$

$$= \sigma_i^2 + \|V_i^T x\|^2 \ .$$

Hence

$$\|\tilde{V}_i^T x\|^2 \geq \sigma_i^2 - \tilde{\sigma}_i^2 \quad . \tag{5.3}$$

Combining (5.1), (5.2), and (5.3) gives

$$\|a\|^2 \geq \frac{(\sigma_i/\tilde{\sigma}_i)^2 - 1}{(\sigma_i/\tilde{\sigma}_i)^2 + 1} \quad .$$

Thus a large value of $(\sigma_i/\tilde{\sigma}_i)^2$ will be reflected by the nearness of $\|a\|^2$ to unity.

6. Acknowledgment

I would like to thank Dr. Michael Saunders, whose good advice introduced me to the algorithm and inspired me to analyze it.

References

1. A. P. Dempster, Elements of Continuous Multivariate Analysis, Addison-Wesley, Reading, Massachusetts (1969).

2. P. E. Gill, G. H. Golub, W. Murray, and M. A. Saunders, Methods for modifying matrix factorizations, Math. Comp. 28, 505-535 (1974).

3. G. H. Golub, Numerical methods for solving least squares problems, Numer. Math. 7, 206-216 (1965).

4. _____, and G. P. Styan, Numerical computations for univariate linear models, J. Stat. Comput. Simul. 2, 253-274 (1974).

5. C. L. Lawson and R. J. Hanson, Solving Least Squares Problems, Prentice-Hall, Englewood Cliffs, New Jersey (1974).

6. G. W. Stewart, Introduction to Matrix Computations, Academic Press, New York (1973).

7. J. H. Wilkinson, The Algebraic Eigenvalue Problem, Clarendon Press, Oxford (1965).

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ONR-N00014-76-C-0391-582	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): THE EFFECTS OF ROUNDING ERROR ON AN ALGORITHM FOR DOWNDATING A CHOLESKY FACTORIZATION.	5. TYPE OF REPORT & PERIOD COVERED: Technical Report.
	6. PERFORMING ORG. REPORT NUMBER: TR-582
7. AUTHOR(s): G. W. Stewart	8. CONTRACT OR GRANT NUMBER(s): N00014-76-C-0391
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Computer Science, University of Maryland, College Park, MD 20742	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Mathematics Branch, Office of Naval Research, Arlington, VA 22217	12. REPORT DATE: Sep[illegible]77
	13. NUMBER OF PAGES: 24
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Cholesky factorization
updating
downdating
regression
least squares
QR factorization

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Let the positive definite matrix A have a Cholesky factorization $A = R^T R$. For a given vector x suppose that $A' = A - xx^T$ has a Cholesky factorization $A' = R'^T R'$. This paper considers an algorithm for computing R' from R and x and an extension for removing a row from the QR factorization of a regression problem. It is shown that the algorithm is stable in the presence of rounding errors. However, it is also shown that the matrix R' can be a very ill-conditioned function of R and x.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED 409 022

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)